Ukraine Conflict Similar Tweets

Gabriele Cerizza

Università degli Studi di Milano
gabriele.cerizza@studenti.unimi.it
https://github.com/gabrielecerizza/amd_project

1 Introduction

In this report we detail our findings in the study of an algorithm capable of identifying similar pairs of documents within massive datasets. The experiments illustrated hereinafter were carried out as part of a project for the Algorithms for Massive Datasets course of Università degli Studi di Milano.

2 Dataset

The dataset employed in our experiments was the "Ukraine Conflict Twitter Dataset" from Kaggle[1], released under the CC BY-NC-SA 4.0 license. This dataset boasts a total of over 40 million tweets concerning the conflict between Russia and Ukraine, which broke out on the 24$^{\text{th}}$ of February 2022. Those tweets were collected daily since the same date by monitoring hashtags. The dataset was first published on the 27$^{\text{th}}$ of February 2022. In the remainder of this report, we will refer to the version 127 of the dataset, downloaded on the 19$^{\text{th}}$ of June 2022.

The dataset comprises 109 compressed CSV files. For each sampled tweet, we can find information concerning the author, the text, the hashtags, the language and the date of creation of the messages. It is worth noting that the naming of the files is not consistent, which hinders attempts to process the tweets chronologically. (Image of a tweet in dataframe?)

2.1 Preprocessing

A number of preprocessing steps were performed on the text of the tweets, having two objectives in mind: (i) discarding noisy and irrelevant information; and (ii) reducing the number of different characters that could be found in the tweets, which adversely affects the model complexity (see further ...). Additionally, we dropped duplicates within each given file, having found a sizeable number of

In particular, we performed the following

might skew

each of which organized in fields providing information about the author each of which

foregoing

Not actually deduped.

[1] www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows

2.2 Conclusion

With regard to the paragraph classification task, our MTL model outperforms the baseline methods. Further improvement can be attained with a more fine-grained analysis of the Wikidata properties of each Wikipedia page.

Our event model struggles to correctly identify trigger words. One reason is that the vast majority of the spans do not contain events and, therefore, finding the correct event span and the correct event type becomes a highly imbalanced problem. The EventGen model fares better and the predictions show that in many cases the mistakes were semantically close to the gold labels. Our argument model compares favourably with the baseline methods. We could improve the model by appending argument type constraints to the templates.

References